# DeepPhysio: Monitored Physiotherapeutic Exercise in the Comfort of your Own Home

Gianmarco Sanesi, Andrew D. Bagdanov, Marco Bertini, and Alberto Del Bimbo
gianmarco.sanesi@stud.unifi.it
firstname.lastname@unifi.it
Media Integration and Communication Center, University of Florence
Firenze, Italy

## ABSTRACT

This paper describes an action classification pipeline for detecting and evaluating correct execution of actions in video recorded by smartphone cameras; the use case is that of simplifying monitoring of how physiotherapeutic exercises are performed by patients in the comfort of their own home, reducing the need of physical presence of therapists. Our approach is based on applying DensePose to every frame of acquired video and subsequent sequence analysis by an LSTM network. We validate our proposed recognition approach on a subset of the NTU RGB+D dataset in order to determine the best classification pipeline for this application. We also describe a mobile, cross-platform application called DeepPhysio that is designed to allow at physiotherapy patients to obtain immediate feedback about the correctness of the physical exercises. Preliminary usability analysis shows that this type of application can be effective at monitoring physiotherapy exercises.

## CCS CONCEPTS

• **Human-centered computing → Interaction paradigms**; *HCI design and evaluation methods*; *Interaction techniques*; • **Computing methodologies → Activity recognition and understanding**; **Neural networks**.

## KEYWORDS

Computer vision, action recognition, medical applications, mobile applications

## 1 INTRODUCTION

The correct execution of instructions received by patients following medical procedures is crucial for the success of therapeutic activity, however the patient's understanding of these instructions is often problematic, whether due to the intrinsic complexity of the actions that must be executed or due to the linguistic and cognitive status of the patients themselves. Indeed, everyday language can be very ambiguous and the more activities must be precisely defined for their effectiveness the more language complexity grows.

Moreover this activity is both time consuming and frustrating for medical staff that must devote an important part of their time to give instructions which are often repetitive. Furthermore, the monitoring of how instructions are performed and the patient's response to treatment is a costly process for all parties involved (patients, family, the health care system), and is a frequently neglected part of the therapeutic process for this reason.

Considering the specific case of physiotherapy, it is has to be noted that physiotherapy is expensive and time consuming, requires clinical visits, exercises prescribed but with no verification of correct performance. A solution to reduce this costly process is to provide a system that simplifies the communication of the therapist with the patient and allow some form of automatized and remotized control of the patient. This is possible thanks to the recent advances in computer vision, especially in action and activity recognition, that might offer automatic ways to monitor performance of physiotherapeutic exercises in the comfort of their own home.

In this paper we describe the DeepPhysio application prototype, developed within the IMAGACT-MED[1] project, that envisions the development and testing of a system prototype that gives e-health support to instructing and monitoring activities. We develop techniques for activity detection capable of determining if a patient is correctly performing a prescribed action.

## 2 SYSTEM DESIGN

*Action recognition.* In our action recognition pipeline, we use DensePose [1] to extract the keypoints of the human figure performing a physical exercise in a video; we selected 17 2-D keypoints of the body, discarding those localized on the head. We trained the pipeline using the dataset NTU RGB+D [2], from which we selected 10 classes of actions in which the subject generally moves similarly to physiotherapeutic exercises (Pick Up, Throw, Sitting Down, Standing Up, Hand Waving, Kicking Something, Hopping, Jump Up, Pointing to Something, Bow).

In order to detect only one bounding box around the subject, we evaluated three different procedures: *i)* Best Score: in each frame of the video we select the bounding box that has the best confidence score of having inside a human being; *ii)* Best Area: we select the bounding box that has the best area among those detected by the ROI proposal system; *iii)* Best Score & Area: we select the bounding box that has the best area among those that have a confidence score greater than or equal to 0.7.

We have also evaluated two different procedures for updating video keypoints' coordinates, collected in $[n\_frames, 17, 2]$ lists: *i)* single barycenter: all keypoint positions are updated relatively to

[1] https://www.micc.unifi.it/projects/imagact-med/

*MM '19, October 21–25, 2019, Nice, France*

ACM ISBN 978-1-4503-6889-6/19/10.
https://doi.org/10.1145/3343031.3350605

the midpoint's position of their frames' geometric barycenters; *ii)* frames barycenter: in each frame keypoint positions are updated relatively to their geometric barycenter.

We use an LSTM network in the our action recognition pipeline to recognize, from the $[n\_frames, 17, 2]$ keypoint position lists, the action performed by the subject in each video. After a list is transformed to $[n\_frames, 12, 2]$, the pipeline detects if it is valid by counting how many frames a keypoint is missing ((None, None) position). If there is one of these that is missing for one third of the video's frames, the list is considered invalid. If a list is valid, it is later analyzed in order to resolve any (None, None) position in the sequence, and finally it is divided in chunks of 29 frames (the minimum length of actions in the dataset) using the sliding window technique. Using an LSTM, in order to predict an action from these chunks obtained from a valid list, we evaluated two different procedures: *i)* max probability prediction: the predicted action for a list is the activity that was detected with the best probability through the analysis of all the chunks; *ii)* voting prediction: the predicted action is the activity that was detected more times through the analysis of all the chunks. The network ha 32 neurons in the recurrent LSTM layer, a final fully connect softmax layer with categorical loss entropy and with RMSprop as optimizer. Training has been done on 200 epochs with batch size of 128 elements.

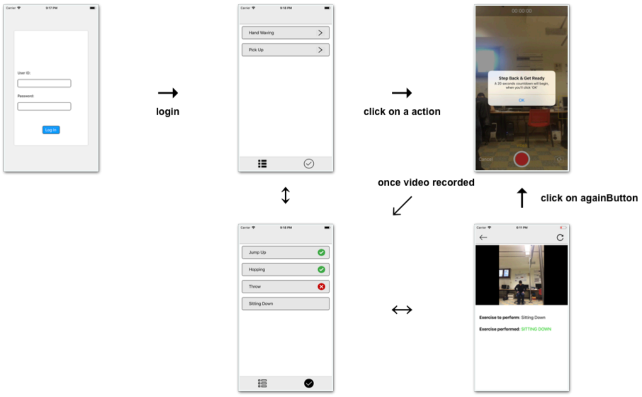

**Figure 1: Workflow of the DeepPhysio app: log in and select the assigned activities. Look at an instruction video and record the activity. Recorded video is sent to the server that evaluates the execution, providing a 3-levels score.**

*The DeepPhysio Application.* The DeepPhysio application was developed to give patients the opportunity to use our action recognition pipeline for the recognition and the detection of the correctness of physical exercise execution. It is mainly composed of:

- a cross-platform mobile application (client-side), developed in Appcelerator Titanium, that allows users to perform their action in front of the smartphone's camera and to observe action recognition pipeline's results;
- a server-side application, developed in Flask, that have to receive the videos recorded through the app, analyze them through the action recognition pipeline and recognize the action performed;
- a real-time client-server communication using Socket.IO.

Using the application the patient can check the list of activities that were prescribed to him, seeing explanatory videos that show how to perform the activities. Then he can record using his own mobile phone a video while performing the activity and receive a feedback about correct, wrong or "uncertain" (i.e. not completely correct, nor wrong) execution; a user can check again the video of how he performed the action and compare it to the given video instruction. Based on this feedback he can retry to perform the activity until he is able to perform it correctly. Usability tests have shown that users that performed wrongly or with some uncertainty an activity were able to perform it correctly after receiving this type of feedback.

| | Max Prob | Voting |
|---|---|---|
| **Best Score - single bar.** | 87.04% | 87.84% |
| **Best Score - frames bar.** | 84.73% | 85.07% |
| **Best Area - single bar.** | 86.56% | 87.23% |
| **Best Area - frames bar.** | 83.99% | 84.38% |
| **Best Score & Area - single bar.** | 87.29% | **87.88%** |
| **Best Score & Area - frames bar.** | 84.94% | 85.34% |

**Table 1: LSTM's results (accuracy) on the created datasets per prediction procedure.**

## 3 EXPERIMENTAL RESULTS

Table 1 reports the results of the LSTM's analysis on the created datasets, composed of 316 sequences per action, divided in 1760 videos for training, 450 for validation and the remaining 950 for testing. The bounding box selection technique *Best Score & Area* has determined the best performances; this is due to the fact that in NTU dataset's videos, not always a single person is shot, as it may also happen when the system is used by a patient at home. Regarding the procedures for updating the keypoints' positions according to the center of the executed action, *single barycenter* technique has resulted to be the most appropriate, as well as *voting prediction* has resulted to be the best action prediction procedure. The best performance obtained using the same visual features and using an SVM as classifier has resulted in an accuracy of 57.52%.

During the usability test, we observed the classifier's behaviour on videos recorded by testers through the application. This analysis has been essentially divided into two phases: firstly, we have examined LSTM's results on keypoints' sequences extracted from videos that were recorded by users on their first try of an action (63% correctly performed actions) and, secondly, we also have examined LSTM's predictions on videos of actions that were recorded another time by testers, using the specific feature in DeepPhysio, when the first execution was not deemed correct, resulting in a 86% correct detection. This shows that users become more aware about how to perform the activity and are able to improve its execution.

## ACKNOWLEDGMENTS

This work was partially supported by Regione Toscana PAR FAS 2007-2013 IMAGACT-MED project.

## REFERENCES

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[2] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. [n. d.]. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*